



METHODOLOGY

Open Access

# Imputation method for lifetime exposure assessment in air pollution epidemiologic studies

Jan Beyea<sup>1\*</sup>, Steven D Stellman<sup>2</sup>, Susan Teitelbaum<sup>3</sup>, Irina Mordukhovich<sup>4</sup> and Marilie D Gammon<sup>4</sup>

## Abstract

**Background:** Environmental epidemiology, when focused on the life course of exposure to a specific pollutant, requires historical exposure estimates that are difficult to obtain for the full time period due to gaps in the historical record, especially in earlier years. We show that these gaps can be filled by applying multiple imputation methods to a formal risk equation that incorporates lifetime exposure. We also address challenges that arise, including choice of imputation method, potential bias in regression coefficients, and uncertainty in age-at-exposure sensitivities.

**Methods:** During time periods when parameters needed in the risk equation are missing for an individual, the parameters are filled by an imputation model using group level information or interpolation. A random component is added to match the variance found in the estimates for study subjects not needing imputation. The process is repeated to obtain multiple data sets, whose regressions against health data can be combined statistically to develop confidence limits using Rubin's rules to account for the uncertainty introduced by the imputations. To test for possible recall bias between cases and controls, which can occur when historical residence location is obtained by interview, and which can lead to misclassification of imputed exposure by disease status, we introduce an "incompleteness index," equal to the percentage of dose imputed (PDI) for a subject. "Effective doses" can be computed using different functional dependencies of relative risk on age of exposure, allowing intercomparison of different risk models. To illustrate our approach, we quantify lifetime exposure (dose) from traffic air pollution in an established case-control study on Long Island, New York, where considerable in-migration occurred over a period of many decades.

**Results:** The major result is the described approach to imputation. The illustrative example revealed potential recall bias, suggesting that regressions against health data should be done as a function of PDI to check for consistency of results. The 1% of study subjects who lived for long durations near heavily trafficked intersections, had very high cumulative exposures. Thus, imputation methods must be designed to reproduce non-standard distributions.

**Conclusions:** Our approach meets a number of methodological challenges to extending historical exposure reconstruction over a lifetime and shows promise for environmental epidemiology. Application to assessment of breast cancer risks will be reported in a subsequent manuscript.

**Keywords:** Exposure, Air pollution, Traffic, Benzo(a)pyrene, PAH, Multiple imputation, Epidemiology, In-migration, Dose

\* Correspondence: [jbeyea@cipi.com](mailto:jbeyea@cipi.com)

<sup>1</sup>Consulting in the Public Interest (CIPi), Lambertville, NJ

Full list of author information is available at the end of the article

## Background

Historical reconstruction of exposure offers the opportunity to cover periods of a person's life when they may be most vulnerable [1,2], but data challenges may make it difficult to attempt such a reconstruction for an epidemiologic study. Despite these difficulties, long-term assessment of residential exposure to pollutants is increasingly being used in environmental epidemiology of disease, particularly cancer [3-5]. As part of an effort in a large-scale study of environmental exposures and breast cancer [6], we wanted to add individualized inhalation exposures of PAH from traffic emissions to a previously collected set of individualized PAH exposures from diet and active/passive smoking. Within the category of PAH air exposure, we focused on traffic emissions because they are a major source of both indoor and outdoor exposures to PAH, and often the largest source in areas near cities, as has been confirmed in a number of experimental studies [7-12]. However, it was necessary to confront a number of methodological challenges before inhalation exposure from traffic could be obtained over a lifetime and before it could be totaled appropriately.

First, it was necessary to decide how to combine yearly exposures over decades. There was no guarantee that simply adding them to obtain a dose would capture possible differences in relative risk per unit of exposure at different study-subject ages. We approached this challenge by computing "effective doses" assuming a range of different functional dependencies on age of exposure. Effective doses are computed by summing functions of yearly exposure. Having a suite of effective doses available for use in regressions against health outcomes allows for broad coverage of the possible age-sensitivities.

Second, it was necessary to deal with exposure before study subjects arrived in the study area, as well as exposure that occurred in years before desirable records on emissions and traffic flows were available. A third challenge arose, because the earlier in time we went back to estimate exposure, the greater the percentage we found of missing residential information. To deal with these two types of missing-data problems we propose the use of multiple imputation, namely constructing a series of data sets with missing parameters randomly drawn from a set of predetermined values derived from an imputation model. The resulting multiple data sets are then available to be regressed against health outcomes. On the positive side of this approach, random errors in regression coefficients are reduced within an imputation data set, because sample sizes are increased. On the negative side, there is new variance in regression coefficients introduced when comparing results between imputed data sets. The net effect of these competing tendencies on overall variance can be assessed using

Rubin's Rules or alternate formulations [13-15]. With this combined variance in hand, confidence limits on regression coefficients can be obtained that explicitly include uncertainty introduced by imputation of missing values.

The fourth challenge we faced was a consequence of limitations in the imputation process itself, namely the vulnerability to increased risk of recall bias, when information needed to impute a missing dose component is based in part on questionnaire answers obtained from individual study participants. This can lead to bias in regression coefficients and, therefore, the possibility of finding false or hidden associations with health outcome.

The quantity imputed is the portion of the total dose accumulated during periods of missing location data or missing emissions data. The percentage of the missing component can vary in magnitude by study subject. Although dealing with partially missing information presented a challenge, it also offered an opportunity to check the robustness of regressions with imputed data, and thus test for recall bias. To this end, we introduce the "percentage of dose imputed" (PDI) as a continuous index that allows limiting analysis to groups of study subjects depending on their degree of imputation. With PDI limit set to 0, the analysis is a complete-case analysis. With the limit on PDI set to 100%, the full population is analyzed. In this way, regression of health data could be performed as a function of the limit on PDI to see if the results were consistent as larger and larger components of imputed dose were allowed for study subjects. If there was consistency, then it would be reasonable to extend the results of a complete-case analysis to larger segments of the full population.

A fifth methodological challenge arose, because the observed doses among the subjects with complete information turned out to be right skewed: specifically in our case-control study the 1% of women who had lived for a long time near heavily trafficked intersections. This presented a challenge to our imputing doses in such a way as to properly populate "outlier" values. To deal with this type of imputation challenge, we propose approaches that do not assume standard distributions; moreover we recommend using more than one approach to allow testing the sensitivity of the regression results to the choice of imputation model.

All of these challenges are likely to be faced, to a lesser or greater degree, when researchers contemplate studies that require individualized historical exposures of adults over a lifetime, or even a great portion of it. As a result of all the potential difficulties, historical exposure is often ignored and analysts look for associations with exposures in recent periods when data are complete. However, doing so can introduce exposure misclassification,

depending on the historical mobility of the population and the variation of exposure in place and time, with a resulting tendency to bias results towards the null. Associations can be missed [16]; and periods of greatest importance for health effects go unnoticed [4]. One approach to dealing with missing historical exposure data is to restrict analysis to a subset of the study population with complete data, but such a restriction can decrease sample size drastically and raises questions about applicability of results to the entire study population. Thus, it was of interest to explore new ways to accomplish multiple imputation in the historical context and to find ways to test for recall bias.

We illustrate the approaches developed to deal with the challenges using data from a population-based, case control study of breast cancer on Long Island, New York, a cancer for which a life-course approach to research is strongly recommended [17].

## Methods

### Risk equation

Our ultimate goal is to obtain dose estimates that can be used as predictor variables in logistic regression models of disease risk in which log (odds of disease) is linear in integrated exposure opportunity (dose).<sup>a</sup>

$$\ln\left(\frac{p_i}{1-p_i}\right) = A + \beta * [Z_{calc}(i, y) + Z_{imp}(i, y)] \quad (1)$$

In Eq. 1,  $p_i$  is the probability of an individual,  $i$ , becoming a case in year,  $y$ .  $\beta$  is a regression coefficient.  $Z_{calc}(i, y)$  is the exposure opportunity accumulated for time periods when information is complete for individuals,  $i$ , permitting individualized estimates.  $Z_{imp}(i, y)$ , is an imputed dose for any time periods of interest when information needed for an individualized exposure estimate is lacking. Only the missing subcomponents of the exposure model are imputed, so that as much physical and biologic knowledge is retained.  $Z_{imp}(i, y)$ , by definition, will be zero for the set of subjects included in complete-case analysis.

Covariates that represent known risk factors for the disease under study, as well as possible confounding variables, are implicitly included in  $A$ . The magnitude of  $Z_{calc}(i, y) + Z_{imp}(i, y)$  determines the risk that subject,  $i$ , will become a case in year,  $y$ , arising from exposure terminating in a given year, such as a year related to a biological state (e.g., menopause), or at the end of the study period, which is the value chosen for our numerical examples.

To account for different sensitivity to age of exposure, the  $Z(i, y)$  terms in Equation 2 allow for more than a simple sum over short-term exposure calculations. Within the sum, each short-term calculation is multiplied

by a “biological effectiveness” factor potentially dependent on age at exposure and age of disease onset.

$$Z(i, y) = \sum_{y'=y_b}^{y_{stop}} E(y', i) * B(y'-y_b, y-y_b), \quad (2)$$

$E(y', i)$  is the short-term exposure at date,  $y'$ . The variable,  $y$ , is the year of disease diagnosis. The sum in Eq. 2 is over the dates,  $y'$ , running from birth year,  $y_b$ , to the last date of exposure interest,  $y_{stop}$ .  $B(y'-y_b, y-y_b)$  is a relative risk that compares the effect of unit exposure at age,  $y'-y_b$ , to the effect at a standard age of exposure, say 1-years old. Thus, if a unit dose at age 10 will double the risk of disease caused by a unit dose at age 1, the  $B$ -variable for age 10 equals 2. In this way, the  $Z$ -terms, which depend on  $B$ , can be called effective doses.

Subjects with the same  $Z$ -dose have the same increase in log (odds of disease) compared to an unexposed person, no matter how their exposure was distributed over time. Rather than define the baseline of  $B$  as a single exposure in a particular year, the baseline can be defined relative to a standard distribution of exposure over a time period. To include birth-cohort effects, for instance in a cohort study, the  $B$ -factor would include a separate entry for birth year,  $y_b$ , and be written as  $B(y'-y_b, y-y_b, y_b)$ .

In this paper, we consider four functional forms for  $B$ , distinguished by subscripts 1 through 4, which span a wide range of possible age sensitivities:

$B_1$ :  $B = \text{Constant}$ ;

$B_2$ :  $B$  is inversely proportional to square of age of onset, which increases relative risk at young ages (as in best mathematical fit to data for relative risk of breast cancer following ionizing radiation exposure [2]);

$B_3$ :  $B$  vanishes except at peak historical exposure (threshold model). (To work as a threshold model,  $B_3$  can be used in conjunction with a threshold step function added to the risk equation.);

$B_4$ :  $B$  is negligible except in the last  $k$ -years of the study (pure promoter model).

We note that a latency period can be implemented in any of the above models simply by changing the stop year in the summations.

The equation for effective dose, Eq. 2, involves a summation over a study subject's entire life, but individual values of each variable that are needed to compute the sum are not usually completely available since birth. To fill in the missing information, we replace variables at various times with group level values. The time steps over which emissions must be tracked in equation 2 depends on the model being used to assign individualized exposure estimates. The base unit might be months, or

any duration corresponding to data typically available depending on disease and the nature of the short-term exposure model. For instance, the natural starting unit might be hours for a meteorological dispersion model that tracks emissions that vary by hour of day and accounts for the correlated meteorological conditions that also change hour by hour, such as rainfall, wind speed, and wind direction. Such an hourly model may be testable against hourly air concentration data collected at monitoring stations [18,19]. As long as average emissions per month do not change rapidly, it possible to do the hourly summations once for each month of the year and use the results for the same month in each year without much loss in accuracy. Data averaged by month is appropriate when movement of study subjects between residences is known by month of the year.

It is useful to separate the exposure,  $E(y', i)$ , in Eq. 2 into sums or products of terms that reflect the underlying model of exposure, so that terms needing imputation can be separately identified. For a land use regression model the natural separation might be a linear combination of terms, including air concentrations at monitoring stations [20]. For a meteorological dispersion model, it would be natural to break  $E(y', i)$  into a product of terms. Regardless of whether the separation is a sum or product, whenever there is a time that a particular term or subpart is missing for an individual, it will need to be imputed according to an imputation model. In either case, random numbers are used in the multiple imputation model as part of the process of drawing parameter values from a predetermined set of options.

For the traffic emission example, we assume organ doses are linearly related to air concentrations and expand Eq. 2 as a product that reflects the underlying physical process of tailpipe carcinogens being emitted from vehicles and those carcinogens being transported through the air to reach a residence:

$$E(y', i) = \varepsilon(y') * T(y') * D(y', r_i[y']), \quad (3)$$

$\varepsilon(y')$  in Eq. 3 is the average emission per km from the tailpipe of an average vehicle along the road network at a particular date,  $y'$ , determined from average US values [21].  $T(y')$ , in Eq. 3, is the average number of vehicles per km on an average road in the modeling area in the year,  $y'$ . Since 1970 it has increased by a factor of 2.7 in the entire U.S. [21].  $D(y', r_i[y'])$ , in Eq. 3, is the “dispersion function,” also called the “transfer function” [22]. It relates emissions from the entire road network to exposure at individual residences,  $r_i$ . It is computed by tracking puffs of pollution as they change direction with hourly-wind changes, as they expand at rates dependent on meteorological conditions, as their concentration

declines due to depletion processes, and as they finally arrive at residences.

$D$  represents one hour’s exposure received at the residence of a subject,  $r_i[y']$  at a date,  $y'$ , from one unit of pollutant emitted from all the vehicles on the road network, some of whose emissions may have started their airborne travel several hours earlier. To complete the calculation of  $D$ , the received exposure is divided by the average number of cars per km on the roadway,  $T(y')$ . In this paper, any imputation of Eq. 3 that is needed for a study subject in year,  $y'$ , is done on the transfer function,  $D(y')$ , as long as data for  $\varepsilon()$  and  $T()$  are available.  $D$ -values tend to be roughly constant over historical time, changing only as the relative traffic pattern changes in different parts of the modeling region. In addition to depending on the road network,  $D$ -values depend on region-specific meteorological factors that enter the transport dispersion model.

#### Incompleteness index

As previously stated, due to incompleteness in reported residence addresses and gaps in available historical exposure data, most epidemiologic studies cannot be expected to have a complete, individualized lifetime dose for every subject. We wish to assess the potential impact on estimates of associations between exposure and effect measures of filling these gaps with imputed values. To this end we define in Equation 4 an, “incompleteness index,” as the group-level imputed dose expressed as a percentage of the total dose:

$$\begin{aligned} \text{Incompleteness} &= \text{PDI} \\ &= \frac{100\% * (\text{imputed dose component})}{\text{imputed dose component} + \text{modeled dose component}} \\ &= \frac{100 * Z_{\text{imp}}}{(Z_{\text{imp}} + Z_{\text{calc}})}, \end{aligned} \quad (4)$$

Our approach is to use this incompleteness index (PDI) to define subsets of participants with varying contributions of group level components. Thus, we might compare regression or correlation results for a subset of study subjects with PDI’s = 0 to larger subsets with PDI’s less than 20, 30, 40%, and so on. If the results are robust, with means and confidence limits that suggest a similar association, we can have confidence that imputation is not introducing artificial findings. Extrapolation of the results to the full sample of study subjects would become reasonable. Chi-square analysis of Tables of cases and controls grouped by intervals of PDI is a simple way to test for differential bias in imputation percentage.

It would also be reasonable to define an alternate incompleteness index based on the percentage of time that imputation was needed for a subject. However, this



“percentage of time imputed” (PTI) is not as conservative an index as the percentage of dose imputed, particularly in situations where historical exposures varied greatly over time. A major advantage of setting a numerical value of PDI as a limit is that it insures that all study subjects have their degree of imputation constrained. In any case, PTI should be computed along with the PDI as a sensitivity check.

#### Imputation of gaps in an individual's dose

For the type of study we are considering, there are boundaries in space and time separating regions where individualized exposure estimates can be made. Yet, some exposure is likely to have occurred outside these boundaries, necessitating group level surrogates for these regions. Surrogates should be consistent at the boundaries such that averages and variances of exposure match the comparable quantities for subjects with individualized exposure. For instance, consider a situation where reliable local traffic and emission data were not available before 1960. Suppose that the number of registered vehicles in a state is taken as the group level surrogate for exposure prior to 1960. For consistency, when the surrogate is extended to a year when it is not needed, e.g., 1960, it should have the same average as the individualized exposures that were computed for 1960. This consistency can be accomplished by scaling the surrogate by the average individualized exposure in 1960,  $dose(1960)$ , and dividing by the vehicle registration number in 1960, as in Equation 5.

$$EarlySurrogate(t, i) = V_1(i) * \left( \frac{Vehicles(t)}{Vehicles(1960)} \right)^{k_1} dose(1960), \quad (5)$$

To individualize the surrogate in Eq. 5 for study subject,  $i$ , the surrogate has been multiplied by a random variable,  $V_1(i)$ , which is uniquely chosen for each study subject in each imputed data set. The distribution of  $V_1$  should be chosen to have a mean of unity and a variance that matches the variance in 1960 of the individualized exposure estimates. For purposes of allowing sensitivity analysis, we have added a parameter,  $k_1$ , as an exponent in Eq. 5, which could be varied around unity, should one want to test regressions with health data for different assumptions about the relationship between early vehicle registration numbers and traffic exposure. Standardization of the surrogate is carried out separately for cases and controls. Although Eq. 5 is independent of geographic locations of residences, it would also be possible to perform the analysis separately by geographical subregions of the study area.

A surrogate like the one defined in Equation 5 is the very opposite of an individualized dose, which implies

considerable possible misclassification. And, the surrogate is likely to be highly correlated with age, requiring careful control of the age-variable to avoid residual confounding. On the other hand, if exposure is particularly important at young ages, then exposures in the period requiring surrogates is likely to be important. Also, it is quite possible that increased cancer incidence with age is due in part to the longer time older people have to accumulate exposure and to develop cancer [23]. A middle ground position would be to compare regression coefficients with and without the early dose surrogate included. When included, the implicit assumption is that early dose effects are so strong biologically that an exposure estimate grossly misclassified at the individual level will still be adequate to improve the overall dose classification, thereby facilitating the detection of any effect. Excluding the early dose surrogate is equivalent to assuming a biologic model that has cancer risk from the exposure falling off in importance years after exposure.

The early surrogate cannot contribute directly to any recall bias, because no information from interview plays a role in its initial calculation. Recall bias could nevertheless carry over to the early surrogate, however, because the scaling factor which is computed separately by cases and controls is based on post-1960 individualized data that depends on subject recall of address locations. For some of the numerical examples given in this paper, we exclude the early dose surrogate, because including it does not help us elucidate potential recall bias using the PDI, which is a major focus of the paper.

It may not always be possible to find a boundary on which to match a surrogate. This is the case for subjects during periods when they resided within the temporal boundaries of the study, but outside the spatial regions. The natural way to generate dose surrogates for these subjects during their out-of-area periods is to retain as much data needed for individual calculations as applies outside the spatial region, but use average values for non-available parameters. The needed average values can be computed by summing over the individualized values calculable for study subjects inside the spatial study area during the same time period.

There is no a-priori reason for expecting the averages to be the same in and outside the region, so we multiply the results by a common sensitivity parameter,  $k_2$ , which can be varied in regressions with health data to check the importance of the default assumption that the averages are the same. Let,  $F(t, p_1, p_2, p_3, p_4)$ , be an individualized exposure at time,  $t$ , for a subject residing inside the study area, given values of the parameters,  $p_1, p_2, p_3$ , and  $p_4$ . Assume that  $p_3$  and  $p_4$  are unknown for study subjects outside of the area. Then, the surrogate for a time when a subject,  $i$ , is out of the study area, becomes:

$$\begin{aligned} \text{OutOfAreaSurrogate}(t, i) &= k_2 * V_2(t, i) \\ &*F(t, p_1, p_2, \langle p_3 \rangle, \langle p_4 \rangle), \end{aligned} \quad (6)$$

where  $p_3$  and  $p_4$  have been replaced by  $\langle p_3 \rangle$  and  $\langle p_4 \rangle$ , averages over values for other study subjects, namely those subjects residing in the study area at time,  $t$ . Once again, individualization can be carried out by multiplying by a random variable with mean unity, which we call,  $V_2$ . The variance of  $V_2$  would be chosen to match the variance of subjects with individualized doses inside the study area. With this surrogate, the variance,  $V_2$ , can change over time, so that matching needs to be carried out more than once, say, every five years. As with the earlier surrogate, standardization is done separately for cases and controls.

We note that the parameter,  $k_2$ , could also serve as a regression coefficient to be determined as part of the regression with health data. An alternate approach would be to individualize the parameters  $\langle p_3 \rangle$  and  $\langle p_4 \rangle$  separately. See supplementary information for a discussion of these options.

The two group-level surrogates defined above refer to dose accumulated over different time periods in a subject's life. Obtaining an individual's imputed dose requires summing over both, if both are non-zero.

#### Imputation of missing residence information

There may be additional imputation required for a study subject, even within the temporal and spatial boundaries of the main study, for instance, to impute an individualized,  $D$ -value or other indicator of local exposure. In our example, historical street addresses, including house number, were obtained by interview, which meant relying on the memory of study subjects for input to geocoding software. Additional file 1: Figure S1-1 shows the percentage of geocoding success by year of arrival at residence. For 1960, the success rate was 65%; therefore, 35% of the addresses in that year needed imputation.

We consider two ways to carry out the imputation on the theory that having two methods to compare regression results is a safeguard. Consistent results have the potential to increase confidence in the outcome; inconsistent results would be a warning that the imputation approach was problematic.

Within the mathematical function that defines short-term exposure there will be some residence indicator, whether it be proximity to roads or a meteorological transfer function specific to a residence, such as the  $D$ -values discussed earlier. When a quantity like  $D$  is missing, the easiest method of imputation, suggested by Raaschou-Nielsen et al. [5] is to interpolate between values at residences before and after the missing value. To obtain values that differ between different imputed

data sets, we can multiply the interpolated imputed exposure quantity by a random number with a log normal distribution whose variance is obtained from a lognormal fit to the two calculable quantities about which one is interpolating. For each imputed data set, a distinct value is randomly drawn from the derived lognormal distribution. Alternatively, when there is a missing quantity at starting or stopping dates, extrapolation can be used. In such a case, zero variance is assigned to the lognormal distribution and there is, therefore, no difference in interpolated dose component for a study subject in different imputed data sets for the missing residence period.

This method has the advantage that it relies solely on individual information and does not require the use of covariates in helping to choose the imputed dispersion coefficient. It has the disadvantage that, if there is a long tail in the exposure distribution, it can overstate the imputed exposure components when one of those high values is used for imputation.

Our second method relies on the fact that when information on residence location is obtained by questionnaire, it is likely that most residents will be able to remember at least the city, town, or village of the residence, even if they are not able to remember the full address. This means their location can be narrowed; in the US, we have found location could be narrowed to the Census Division Place level 80% of the time in our example study. Thus, it is possible to narrow the imputation to values of the observed data within the corresponding Place. In what follows, we call this, "imputation by place." To handle the situation when a census division cannot be assigned or when the number of either cases or controls in a census Place is less than five (11% of locations in our example study), it is appropriate to fall back on the interpolation method.

With imputation by place, it is good practice to use covariates available for a subject, such as income or education, in picking the imputed values, because they could have influenced the area chosen for a residence. One of the limitations of multiple imputation is that there are situations involving missingness not at random that can introduce bias. This possibility can be reduced by including a large number of covariates in the imputation model, thereby increasing the chances that any variables predictive of missing values are incorporated [15].

Observed values of covariates included in the imputation-by-place algorithm for our breast cancer example were  $D$ , smoking status (passive and active), education, history of fertility problems, history of benign breast disease, family history of breast cancer, combined estrogen receptor and progesterone receptor status among cases, parity, BMI, dietary intake of PAH, lifetime intakes of grilled or barbecued and smoked meats,

lifetime alcohol intake, and age at diagnosis. Case-control status was also included, which at first consideration seems circular; however, including it is a valid part of multiple imputation [24,25]. Otherwise, patterns in the observed data might not be transferred to the unobserved data during imputation.

In contrast to imputation by interpolation, which may overestimate high values of dose, when there is a highly right-skewed distribution, the method of imputation by place may overestimate some low values of dose, should the values in the top tail of the distribution be correlated with the density of road intersections, as would be expected. This contrasting tendency makes use of the two methods in regression of health outcomes a useful test of the robustness of results.

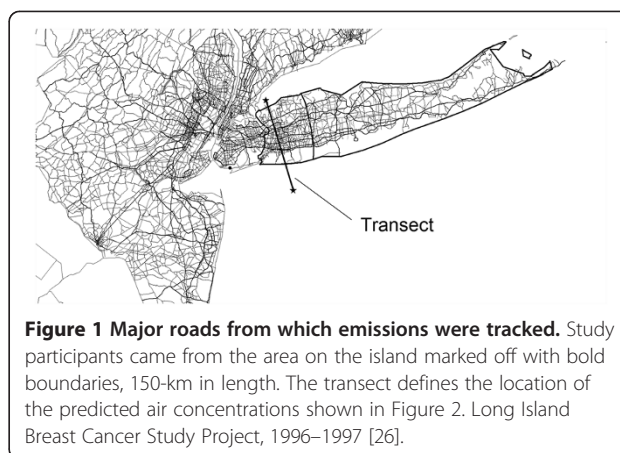
### Breast cancer example

#### Study population

We use data on 3,064 female breast cancer cases and controls who were residents of Long Island, New York, between August, 1996, and July, 1997 [26]. Cases were women newly diagnosed with a first primary *in situ* or invasive breast cancer; controls under 65 were identified through random digit dialing and by Health Care Finance Administration (HCFA) rosters for those 65 years of age and older. Controls were frequency matched to the expected age distribution of cases.

Residence locations on Long Island, but not outside of Long Island, were identified at the time of the in-person, structured interview, starting with a subject's earliest recollection. This resulted in a total of 8,319 locations in Nassau and Suffolk counties divided among the 3,064 women. Address coordinates at the street level were identifiable for 5,383 of these locations using commercial geocoding software (BLR Inc., now part of GDT, Lebanon, NH), with 95% of the remainder identifiable only to the 5-digit Census category, "Place," due to incomplete address information. A census division, "Place," consists of consolidated cities and incorporated places, such as town or village. Certain unincorporated areas may also be assigned a Place number <http://www.census.gov/geo/www/codes/place/download.html>. The average population of a study-area Place in the 1990 Census was approximately ten thousand, or about 1% of the total population of Long Island. Location to the Place level is sufficient to capture the urbanization gradient along the length of the elongated Island study area [18]. For consistency, the street maps used for geocoding were used to model dispersion of traffic pollution. The study area and surrounding traffic network are shown in Figure 1. Relative 1-hour air concentrations predicted by the geographic dispersion model in 1995 along a transect across Long Island are shown in Figure 2.

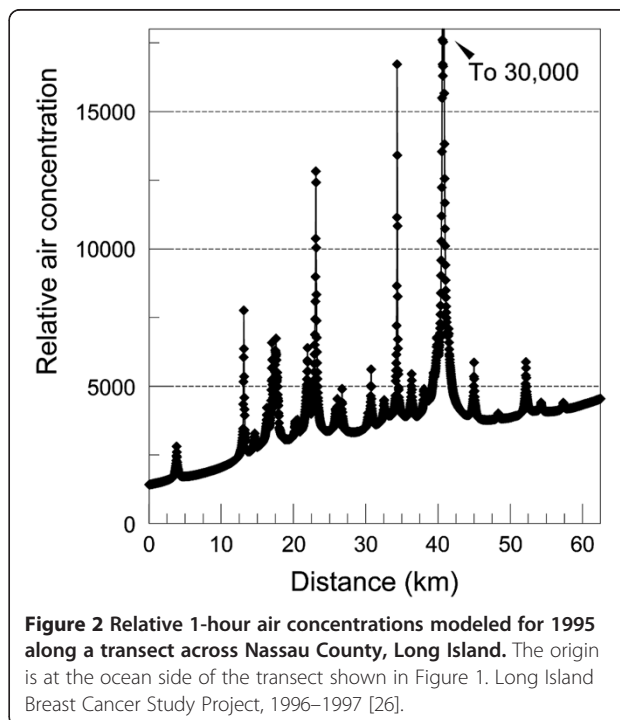
Benzo(a)pyrene (BaP) emitted from vehicle tailpipes was the pollutant proxy analyzed for the numerical



**Figure 1 Major roads from which emissions were tracked.** Study participants came from the area on the island marked off with bold boundaries, 150-km in length. The transect defines the location of the predicted air concentrations shown in Figure 2. Long Island Breast Cancer Study Project, 1996–1997 [26].

examples, because there is an abundance of good historical data [21]. The emission per km variable,  $\epsilon(y')$ , in Eq. 3 peaked in the US in the mid-1970s, declining rapidly with the introduction of catalytic converters until the late 1980s [21]. This means that per unit of time, exposure in the pre-1980 period was likely to be much higher than after 1985, adding large temporal variation to the spatial variation associated with proximity to traffic evident in Figure 2.

Data for BaP and other PAH tailpipe emissions extend back only to 1960, which sets that year as a boundary for shifting to group level information. In the absence of information, and no engineering reason evident in the literature to think otherwise, we took  $\epsilon(y')$  to be a



**Figure 2 Relative 1-hour air concentrations modeled for 1995 along a transect across Nassau County, Long Island.** The origin is at the ocean side of the transect shown in Figure 1. Long Island Breast Cancer Study Project, 1996–1997 [26].

constant prior to 1960 equal to the known value in 1960. We also took the dispersion variable,  $D$ , as constant, equal to the average value computed in 1960 for residences in the study area. We used vehicle registrations in New York State, for which date existed back to 1900, as the surrogate for traffic flow rates for the pre-1960 period, as in Eq. 5. See Additional file 1: Figure S1-2.

#### **The LIBCSP meteorological dispersion model**

The LIBCSP meteorological dispersion model was developed specifically for the Long Island Breast Cancer Study Project's analysis of traffic exposures, although it has been adapted for use in the Buffalo, New York area, with separate tests carried out in that region [19,27]. A special feature of the model is its allowance for separate emissions from road intersections, which are known to be higher than during cruise emissions [28,29]. The model does not account for street canyon effects, but that is not a necessity for Long Island.

The contribution from each road segment to the air concentration at a downwind receptor residence was computed within 100 meters of a road using a highway line-source model [30] applied to each of the 500,000, straight-line road segments in the traffic network. The "Chock" highway model was chosen because it gave the best fit, when compared to a suite of models tested, to tracer concentrations near the Long Island Expressway as part of a test carried out by the New York Department of Environmental Protection [31].  $R^2$  values ranged from 0.75 to 0.92 for various meteorological conditions and angles to the road.

Beyond 100 meters, we used a standard Gaussian puff dispersion model (equivalent to the USEPA's "RAM" model [32]). Additional file 2: Table S2-3 describes the default parameters and data sources used in the dispersion model. Further details can be found in the text of Additional file 2, including an explanation of the approach used to shift between different meteorological models at 100 meters and a description of how we account for background BaP blown into the study area from outside it.

Total concentration at a study subject's residence was computed as the sum of the contributions from the approximately 500,000 source segments. This concentration at a residence forms the basis for the transfer function,  $D$ .

The transport model used to compute the transfer function,  $D$ , was previously validated against BaP soil measurements at approximately 500 residences [18]. A key result was higher than average emissions coming from traffic intersections, which were modeled separately. See Additional file 1: Figure S1-3 for a graphic showing

separately-modeled intersection regions superimposed on a segment of the road network.

#### **Imputation to account for missing residence location in the Long Island example**

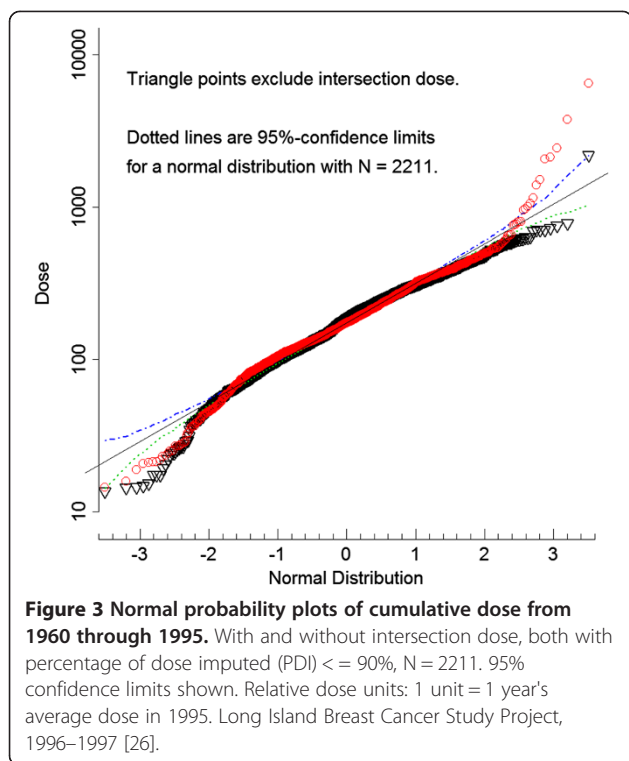
Imputation is needed during times when subjects reside in the study area, but their residence location is not known with enough precision to determine accurate latitudes and longitudes, quantities that are necessary to obtain individualized exposure. In the example data presented in this paper, the primary method of imputation used was imputation by place. We imputed a  $D$ -value from the set of  $D$ 's calculated for other women residing in the same census division Place. The choice of imputed values was conditioned on the observed covariates discussed earlier using fully conditional specification as implemented in the R-program, "Multivariate Imputation by Chained Equations (MICE)" [33]. MICE is but one of a number of imputation programs available in statistics computer packages. Fifteen different sets of data were calculated with missing transfer functions and any missing covariates imputed. For further details of the example imputation, see text in Additional file 2. The input to the imputation program is a matrix file with "NA"s for missing transfer functions and missing covariates; the output is a set of files with the NAs replaced with numeric values, which generally differ from set to set. Because of the long-tailed distribution found in exploratory analysis (see Figure 3), there was no standard distribution, e.g., lognormal, that could be used to fill the missing transfer functions. Therefore, we ran the MICE program with the option to sample randomly from  $D$ -values fully calculable for other subjects. These samplings are not completely random, however; the program uses regression techniques to weight the choices according to any tendency of different transfer functions to cluster by covariates. The output files are then available for regression against health data and the variance of the results between imputed data sets plays a role in determining confidence limits using standard rules [14].

#### **Results**

The main results from this paper are Eqs. 2, 3, 4, 5, 6 and the approaches we described to operationalize them. In this section, we discuss results from the Long Island example.

Different biological effectiveness functions: For the illustrative example, Pearson correlation coefficients between dose calculated with different biological effectiveness factors ranged from 0.33 to 0.79 with no limits set on PDI ( $PDI < = 100\%$ ). The range was 0.59 to 0.99 for complete-case analysis ( $PDI = 0$ ). See Table 1 and Additional file 1: Tables S1-1 and S1-2.





### Early surrogate dosage (before 1960)

The early surrogate dosage for the breast cancer study subjects averaged 50% of the post-1960 dose (calculated for the  $B_1$ , constant sensitivity biologic model). Additional file 1: Figure S1-4 graphs the average dependency on age. The Pearson correlation coefficient was 0.6.

There was essentially zero correlation between the early surrogate and the cumulative dose post-1960. The Pearson correlation coefficient ranged from  $-0.05$  to  $0.001$  as the limit on PDI ranged from 0 to 100%. This low correlation means that inclusion of the early surrogate in the dose summation could alter the distribution curve for total dose.

### Doses calculated after 1960 ( $Z_{calc}$ and $Z_{imp}$ )

Of the 2,986 study subjects who arrived in the study area before 1996, about 55% of them arrived after 1960. Thus, individually modeled doses in the post-arrival period ( $Z_{calc}$ ) are strongly affected by arrival date because late arrivers are missing their out-of-area contribution. Pearson correlation coefficient with year of arrival =  $-0.37$ . Figure 4a graphs the relationship, showing how the  $Z_{calc}$  post-arrival doses fall off for late arriving study subjects. Without accounting for out-of-area exposure through imputation, i.e., by adding the  $Z_{imp}$  contribution, false differences in dose calculations can arise, should cases and controls, because of unmeasured covariates, have a different distribution of arrival times.

The  $Z_{imp}$  surrogate itself, like  $Z_{calc}$ , is highly correlated with year of arrival, albeit with a different sign (Pearson coefficient =  $0.56$ ). Figure 4b shows how adding the post-arrival, out-of-area surrogate,  $Z_{imp}$ , to the individualized dose component,  $Z_{calc}$ , removes the dependence on year of arrival.

To avoid interpreting BaP as more than a PAH proxy [34], doses are presented in the Figures, not in units of inhaled milligrams of BaP, but in units of the 1995 average dose computed for women in the study area. Note that Figures 4a and 4b show graphs for but one imputed data set; however, the patterns are virtually identical at this level of detail for 14 additional imputed data sets, as is to be expected given the great variance around the mean dose curve (results not shown). Also note that a dose value in Figure 4b may be the sum of three terms, an out-of-area imputation, an in-area imputation, and a geographically modeled component.

Although adding  $Z_{calc}$  and  $Z_{imp}$  together eliminates the obvious way that recall bias could enter the estimation of doses, it is still necessary to go further and look for residual bias in  $Z_{calc} + Z_{imp}$ , which can be done using the PDI variable. We performed a chi-square analysis of

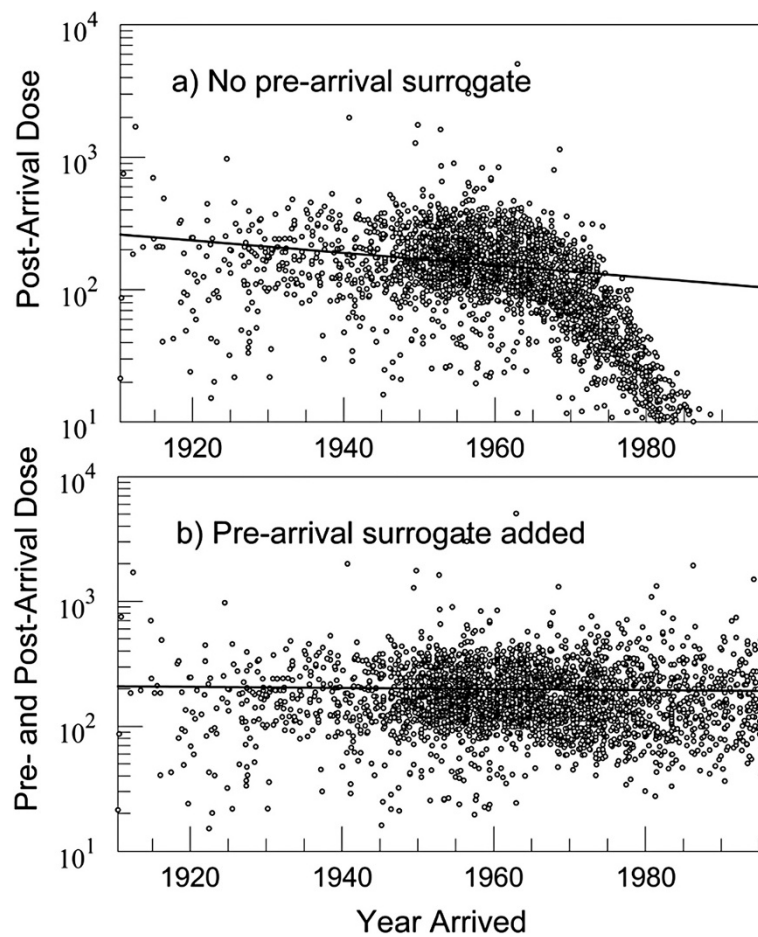
**Table 1 Pearson correlation coefficients between doses computed with different functional forms for the biologic effectiveness factor, averaged over 15 imputations<sup>a</sup>**

Dose surrogate	Samples of women limited by allowed percentage of imputed dose (PDI)						
	CCA (PDI = 0%) N:	PDI < 20% 547	PDI < 40% 1164	PDI < 60% 1477	PDI < 80% 1764	PDI <= 100% 2051	2986
Cumulative dose 1960-1995							(comparison dose variable)
Dose for 1995 only (Promoter model)	0.59 (0.11)	0.39 (0.15)	0.39 (0.16)	0.39 (0.16)	0.41 (0.16)	0.33 (0.15)	
Peak annual dose in 1960–1995 (Threshold model)	0.99 (0.00)	0.93 (0.00)	0.93 (0.00)	0.92 (0.01)	0.91 (0.01)	0.75 (0.017)	
Cumulative dose $\times$ (onset age) <sup>-2</sup> (Age sensitive model) <sup>b</sup>	0.89 (0.024)	0.87 (0.028)	0.85 (0.033)	0.84 (0.039)	0.82 (0.044)	0.79 (0.06)	
Pre-1960 surrogate <sup>c</sup>	$-0.020$	$-0.031$	$-0.031$	$-0.030$	$-0.016$	$+0.00044$	

<sup>a</sup>Numbers in parenthesis are the square root of the variance across imputations divided by the square root of 15, which provides an estimate of the standard deviation of the average over imputations.

<sup>b</sup>Weighting based on the mathematical fit to radiation risks of excess breast cancer in atomic bomb survivors.

<sup>c</sup>Calculated for one imputation only.



**Figure 4 Dose as a function of year of arrival in study area, with/without pre-arrival surrogate.** Relative dose units: 1 unit = 1 year's average dose in 1995. Long Island Breast Cancer Study Project, 1996–1997 [26].

cases and controls grouped by intervals of PDI. The results are shown in Table 2. The p-value for rejecting the null hypothesis was 0.13, which is low enough to suggest some association between PDI and case status in our illustrative example, and therefore the possibility of differential bias.

**Right-skewed distribution and a 1% dose tail**

The top 1% of the cumulative dose distribution showed unusually high values associated with doses from heavily-trafficked intersections, as shown in Figure 3, where probability plots are presented with and without

**Table 2 Distribution of women by limits on incompleteness index (PDI<sup>a</sup>) in cumulative dose for one imputed data set<sup>b</sup> (Women who arrived before 1996)**

Limits on PDI <sup>a</sup>	Number of control women	Number of case women	Total
Zero (complete coverage)	267	280	547
>0 and < 20%	336	281	617
20% to < 40%	164	149	313
40% to < 60%	150	137	287
60% to < 80%	136	151	287
80% to <= 100%	465	470	935
Total	1518	1468	2986
Probability level for Chi-Square			0.13

<sup>a</sup>PDI = Percent of dose imputed.

<sup>b</sup>The Chi-square probability level can vary between low (0.04) and high (0.31) values for different imputed data sets.

the intersection dose component included in total dose. Without intersection dose included, the probability plot of log dose in Figure 3, with the exception of one point, trends smoothly below the normal line at the high end. However, with intersection dose included there is an abrupt deviation, with points falling well outside the upper 95% confidence limit on the normal curve shown in the Figure. The same skewed distribution appears when plotting doses computed for other assumptions about the biological effectiveness factor,  $B$ , with the exception of the pure promoter model,  $B_4$  (results not shown). The skewed distribution at high doses was maintained no matter the limitation placed on imputation percentage (PDI) in 15 imputed data sets (results not shown).

## Discussion

The past decade has seen an increasing understanding of the importance of the impact on chronic disease risk of exposures at all stages of life, especially exposures occurring during the early years of life when many organs are still undergoing rapid and formative growth [17]. Because many epidemiological studies of chronic illnesses such as cancer are usually initiated when the study population have already reached adulthood, capturing decades-old data on critical factors such as nutrition, exercise, and environmental exposures is a major challenge.

Residential address histories can be gathered by interview with relative ease, and we previously developed models for using histories obtained directly from patient interviews to estimate residential exposure to PAHs from traffic sources [35,36] which we then calibrated against environmental measurements for participants in a breast cancer case-control study [18]. We further refined our methodology and historical database to take into account long-term changes in emission control technology [21]. A primary limitation, however, continues to be the completeness and accuracy of patient address recall. The present work provides a methodology to address this type of limitation by applying multiple imputation methods to extract the maximum information from available data, making it possible to use incompletely known residential data to put credible bounds on exposure metrics.

The early (pre-1960) surrogate dosage for the breast cancer study subjects averaged 50% of the post-1960 dose, showing that it can make a substantial contribution to lifetime dosage.<sup>b</sup> Interestingly, because this surrogate prior to scaling does not depend on information obtained from interview, it cannot directly introduce any recall bias into the dose estimates, although some bias can be carried over to it from the post-1960 averages. The way the early surrogate is defined in our example, without separate standardization by geographic region,

the only connection to an individual is through age. However, the weak association ( $p = 0.13$ ) between PDI and case status in our breast cancer data implies possible differential bias due to missing residential data; this will be investigated in a forthcoming epidemiological analysis that uses the present method to predict breast cancer risk.

Our finding of a consistent, strongly right-skewed distribution of traffic doses, associated with heavily trafficked intersections, means that non-standard distributions of exposure must be anticipated in studying traffic emissions. And, for our example data, this high-dose subpopulation is worthy of attention in future risk analyses that include these study subjects.

Because Pearson correlation coefficients between dose calculated with different biological effectiveness factors ranged from 0.33 to 0.79 with no limits set on PDI ( $PDI < = 100$ ), the sensitivity of risk to age at exposure could be important in determining the strength of an association with health outcome; moreover, it suggests that regressions against health data might shed some light on which biologic model is most appropriate.

## Conclusions

In historical reconstructions of exposure, investigators typically identify data that have been collected for environmental monitoring and not for epidemiology, and that do not fully cover the population of interest. We suggest here a way of filling in missing data that we hope will spur increased use of historical exposure reconstruction in health studies. A novel feature of our exposure reconstruction approach is an incompleteness index (PDI) that can be used to study the degree to which imputation might create artifactual dosages and, subsequently, improper associations with health outcomes -- a form of differential bias to which researchers need to be sensitive in general [37,38] and specifically in traffic air pollution studies [39].

## Endnotes

<sup>a</sup>The risk equation in the main text, Eq. 1, was considered to be linear in dose. If the log odds is considered to be non-linear in exposure, then there would be terms in powers of  $Z$  as well a term in the log of  $Z$  [23,40,41] and possibly powers of log of  $Z$ . We did not consider such an extension of the methodology in the examples given in this paper.

<sup>b</sup>One way to reduce the relative contribution of the early surrogate would be to trade one approximation for another, namely running the full exposure model at times earlier than the period for which full emissions and traffic data exists. This extrapolation could introduce misclassification due to errors in traffic flow rates and tailpipe emission rates, but might introduce

less misclassification than would result from relying on an early surrogate. Depending on the variability in a model, an analyst might explore extrapolating the individualized model back in time to reduce the contribution from the early dose surrogate to a modest level relative to total dose.

## Additional files

**Additional file 1: Figure S1-1.** Percentage of addresses successfully geocoded by year of arrival at residence; **Figure S1-2.** Vehicle registrations in New York State by year; **Figure S1-3.** Segments of roads within 100 meters of major intersections where PAH emissions are increased in dispersion model; **Figure S1-4.** Early surrogate (dose before 1960) versus age; **Table S1-1.** Imputation by interpolation: Pearson correlation coefficients between doses computed with different functional forms for the biologic effectiveness factor; **Table S1-2.** Imputation by interpolation: Spearman correlation coefficients between doses computed with different functional forms for the biologic effectiveness factor.

**Additional file 2: Details about imputation of the transfer function, D, using the method of imputation by place. Table S2-1.** Variables imputed and/or used in imputation; Comparison of historical dose quartiles with exposure quartiles at most recent address by dose categories. **Table S2-2.** Chi-Square Table comparing end-of-period (1995) dose to cumulative dose; Issue of multiple comparisons; Additional details of the LIBCSP meteorological dispersion model; **Table S2-3.** Default meteorological dispersion model.

## Abbreviations

A: Constant term in regression, including covariate contributions to risk;  $B_i$  through  $B_4$ : Functional forms for biological effectiveness;  $B(y'-y_b)/y'-y_b$ : Biological effectiveness factor in subject,  $i$ , that relates exposure at a particular biological age,  $y'-y_b$ , to detection of disease at a later biological age,  $y-y_b$ ; BaP: Benzo(a)pyrene;  $\beta$ : Vector of regression coefficients for  $Z(i,y)$ ;  $D(y',y)$ : Transfer function;  $E(y')$ : Exposure at date,  $y'$ , for subject,  $i$ ;  $\epsilon(y')$ : Emission from a vehicle tailpipe in year,  $y'$ ;  $p_i$ : Probability of an individual getting a disease; PDI: Percentage of dose imputed;  $r_i(y')$ : Residence of a subject at date,  $y'$ ;  $T(y')$ : Average number of vehicles on an average road on date  $y'$ ;  $V_1(i)V_2(i)$ : Random variables with mean of unity; Y: Date of disease detection;  $y'$ : Date of an exposure;  $y_b$ : Date of birth;  $y_{stop}$ : Last exposure date of interest;  $Z(i,y)$ : Vector of effective doses for individual,  $i$ , in year,  $y$ ;  $Z_{calc}$ : Modeled component of individual dose;  $Z_{imp}$ : Imputed part of dose.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

JB developed and computed the exposure model; he drafted the manuscript. SS participated in the design and coordination of the approach and helped to draft the manuscript. ST managed study subjects, their interviews, and measurements. IM tested the multiple imputations for utility in regressions with health outcomes, as well as analyzed statistical properties of the exposure distributions. MG conceived of the LIBCSP and the need for a geographic model to assess airborne PAH exposure, managed study subjects, their interviews, and measurements; she promoted the use of multiple imputation and helped to draft the manuscript. All authors read and approved the final manuscript.

## Acknowledgments

Maureen Hatch provided advice during the early stages of this work. In addition, the following individuals provided information and/or assistance: Jing Nie, James Cook, Laura Palmer, Suffolk County Department of Public Works, Michael J. Finocchi, Nassau County Department of Public Works, Ed Finster, USEPA, Roy Fedelem, Long Island Regional Planning Board, Anthony Dentrone, Hagstrom maps, Sureli Nepomuceno, Columbia University Health

Sciences Library. This work was supported in part by Intramural Research Program of NIH, National Cancer Institute, Division of Cancer Epidemiology and Genetics; USPHS grants CA/ES-66572, ES-10126, CA-63021, CA-17613, and, Long Island and Babylon Breast Cancer Coalitions, West Islip Breast Cancer Coalition for L.I., Inc., Huntington Breast Cancer Action Coalition, Noreen T. Holland Breast Cancer Foundation, Inc., Breast Cancer Grassroots Organizations for a Unified Purpose.

## Author details

<sup>1</sup>Consulting in the Public Interest (CIPI), Lambertville, NJ. <sup>2</sup>Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA. <sup>3</sup>Department of Preventive Medicine, Mount Sinai School of Medicine, New York, NY, USA. <sup>4</sup>Department of Epidemiology, University of North Carolina at Chapel Hill, Raleigh, North Carolina, USA.

Received: 26 December 2012 Accepted: 24 July 2013

Published: 7 August 2013

## References

1. Cohn BA, Wolff MS, Cirillo PM, Sholtz RI: DDT and breast cancer in young women: new data on the significance of age at exposure. *Environ Health Perspect* 2007, **115**:1406–1414.
2. Preston DL, Mattsson A, Holmberg E, Shore R, Hildreth NG, Boice JD Jr: Radiation effects on breast cancer risk: a pooled analysis of eight cohorts. *Radiat Res* 2002, **158**:220–235.
3. Lubin JH, Wang ZY, Boice JD Jr, Xu ZY, Blot WJ, Wang LD, Kleinerman RA: Risk of lung cancer and residential radon in China: pooled results of two studies. *Int J Cancer* 2004, **109**:132–137.
4. Nyberg F, Gustavsson P, Jarup L, Bellander T, Berglund N, Jakobsson R, Pershagen G: Urban air pollution and lung cancer in Stockholm. *Epidemiology* 2000, **11**:487–495.
5. Raaschou-Nielsen O, Bak H, Sørensen M, Jensen SS, Ketzel M, Hvidberg M, Schnohr P, Tjønneland A, Overvad K, Loft S: Air pollution from traffic and risk for lung cancer in three danish cohorts. *Cancer Epidemiol Biomarkers Prevent* 2010, **19**:1284–1291.
6. Gammon MD, Santella RM: PAH, genetic susceptibility and breast cancer risk: an update from the long island breast cancer study project. *European J Cancer* 2008, **44**:636–640.
7. Dubowsky SD, Wallace LA, Buckley TJ: The contribution of traffic to indoor concentrations of polycyclic aromatic hydrocarbons. *J Expo Anal Environ Epidemiol* 1999, **9**:312–321.
8. Lim LH, Harrison RM, Harrad S: The contribution of traffic to atmospheric concentrations of polycyclic aromatic hydrocarbons. *Environ Sci Technol* 1999, **33**:3538–3542.
9. Harkov R, Greenberg A, Darack F, Daisey JM, Liroy PJ: Summertime variations in polycyclic aromatic hydrocarbons at four sites in New Jersey. *Environ Sci Technol* 1984, **18**:287–291.
10. Dickhut RM, Canuel EA, Gustafson KE, Liu K, Arzayus KM, Walker SE, Edgcombe G, Gaylor MO, MacDonald EH: Automotive sources of carcinogenic polycyclic aromatic hydrocarbons associated with particulate matter in the Chesapeake Bay Region. *Environ Sci Technol* 2000, **34**:4635–4640.
11. Dunbar JC, Lin CI, Vergucht I, Wong J, Duran JL: Estimating the contributions of mobile sources of PAH to urban air using real-time PAH monitoring. *Sci Total Environ* 2001, **279**:1–19.
12. Levy JI, Houseman EA, Spengler JD, Loh P, Ryan L: Fine particulate matter and polycyclic aromatic hydrocarbon concentration patterns in Roxbury, Massachusetts: a community-based GIS analysis. *Environ Health Perspect* 2001, **109**:341–347.
13. Zhou X, Reiter JP: A note on bayesian inference after multiple imputation. *Am Stat* 2010, **64**:159–163.
14. White IR, Royston P, Wood AM: Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011, **30**:377–399.
15. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR: Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009, **338**:b2393.
16. Alavanja MC, Lubin JH, Mahaffey JA, Brownson RC: Residential radon exposure and risk of lung cancer in Missouri. *Am J Public Health* 1999, **89**:1042–1048.



17. IOM: *Breast Cancer and the Environment: A Life Course Approach*; Committee on Breast Cancer and the Environment, Institute of Medicine. Washington. DC: The National Academies Press; 2012.
18. Beyea J, Hatch M, Stellman SD, Santella RM, Teitelbaum SL, Prokopczyk B, Camann D, Gammon MD: **Validation and calibration of a model used to reconstruct historical exposure to polycyclic aromatic hydrocarbons for use in epidemiologic studies.** *Environ Health Perspect* 2006, **114**:1053–1058.
19. Nie J, Beyea J, Bonner MR, Han D, Vena JE, Rogerson P, Vito D, Muti P, Trevisan M, Edge SB, Freudenheim JL: **Exposure to traffic emissions throughout life and risk of breast cancer: the Western New York Exposures and Breast Cancer (WEB) study.** *Cancer Causes Control* 2007, **18**:947–955.
20. Noth EM, Hammond SK, Biging GS, Tager IB: **A spatial-temporal regression model to predict daily outdoor residential PAH concentrations in an epidemiologic study in Fresno, CA.** *Atmos Environ* 2011, **45**:2394–2403.
21. Beyea J, Stellman SD, Hatch M, Gammon MD: **Airborne Emissions from 1961 to 2004 of Benzo[a]pyrene from U.S. Vehicles per km of Travel Based on Tunnel Studies.** *Environ Sci Technol* 2008, **42**:7315–7320.
22. Liu LJ, Curjuric I, Keidel D, Heldstab J, Kunzli N, Bayer-Oglesby L, Ackermann-Liebrich U, Schindler C: **Characterization of source-specific air pollution exposure for a large population-based Swiss cohort (SAPALDIA).** *Environ Health Perspect* 2007, **115**:1638–1645.
23. Greenland S: **Dose-response and trend analysis in epidemiology: alternatives to categorical analysis.** *Epidemiology* 1995, **6**:356–365.
24. Donders AR, van der Heijden GJ, Stijnen T, Moons KG: **Review: a gentle introduction to imputation of missing values.** *J Clin Epidemiol* 2006, **59**:1087–1091.
25. Little RJA: **Regression With Missing X's: A Review.** *J Am Stat Assoc* 1992, **87**:1227–1237.
26. Gammon MD, Neugut AI, Santella RM, Teitelbaum SL, Britton JA, Terry MB, Eng SM, Wolff MS, Stellman SD, Kabat GC, Levin B, Bradlow HL, Hatch M, Beyea J, Camann D, Trent M, Senie R, Garbowski G, Maffeo C, Montalvan P, Berkowitz G, Kemeny M, Citron M, Schnabel F, Schuss A, Hajdu S, Vinceguerra V, Collman G, O'abrams I: **The long island breast cancer study project: description of a multi-institutional collaboration to identify environmental risk factors for breast cancer.** *Breast Cancer Res Treat* 2002, **74**:235–254.
27. Li Y, Nie J, Beyea J, Rudra CB, Browne RW, Bonner MR, Mu L, Trevisan M, Freudenheim JL: **Exposure to traffic emissions: Associations with biomarkers of antioxidant status and oxidative damage.** *Environ Res* 2013, **121**:31–38.
28. Nelli JP, Messina AD, Bullin JA: **Analysis and modeling of air quality at street intersections.** *JAPCA* 1983, **33**:760–764.
29. Sculley R: **Vehicle emission rate analysis for carbon monoxide hot spot modeling.** *JAPCA* 1989, **39**:1334–1343.
30. Chock DP: **A simple line-source model for dispersion near roadways.** *Atmos Environ* 1978, **12**:823–829.
31. Sistla G, Samson P, Keenan M, Rao S: **A study of pollutant dispersion near highways.** *Atmos Environ* 1979, **13**:669–685.
32. Catalano JA, Turner BD, Novak JH: **User's Guide for RAM.** In *Research Triangle Park*. 2nd edition. Environmental Protection Agency; 1987.
33. van Buuren S, Groothuis-Oudshoorn K: **Multivariate imputation by chained equations in R.** *J Stat Software* 2011, **45**:1–67.
34. Stellman SD, Guidotti TL: **Polycyclic aromatic hydrocarbons.** In *Environmental and Occupational Medicine*. Edited by Rom WN, Markowitz SB3. Philadelphia: Lippincott Williams & Wilkins; 2007:1240–1250.
35. Beyea J, Hatch M: **Geographic exposure modeling: a valuable extension of GIS for use in environmental epidemiology mht version.** *Env Health Perspectives* 1999, **107**(Suppl. 1):181–190.
36. *Development of a traffic model for predicting airborne PAH exposures since 1960 on Long Island.* New York: Report to the National Cancer Institute and the National Institute of Environmental Health Sciences for work completed under USPHS Grant U01-CA/ES-66572. <http://www.cipi.com/PDF/beyea2005traffcpahmodel.pdf>.
37. Chyou P-H: **Patterns of bias due to differential misclassification by case-control status in a case-control study.** *European J Epidemiol* 2007, **22**:7–17.
38. Jurek AM, Greenland S, Maldonado G: **Brief Report: How far from non-differential does exposure or disease misclassification have to be to bias measures of association away from the null?** *Int J Epidemiol* 2008, **37**:382–385.
39. Rosenlund M, Bellander T, Nordquist T, Alfredsson L: **Traffic-generated air pollution and myocardial infarction.** *Epidemiology* 2009, **20**:265–271.
40. Boucher KM, Slattery ML, Berry TD, Quesenberry C, Anderson K: **Statistical methods in epidemiology: a comparison of statistical methods to analyze dose-response and trend analysis in epidemiologic studies.** *J Clin Epidemiol* 1998, **51**:1223–1233.
41. Becher H: **Analysis of continuous covariables in epidemiological studies: Dose-response modelling and confounder adjustment.** *Biometrical J* 2002, **44**:683–700.

doi:10.1186/1476-069X-12-62

**Cite this article as:** Beyea et al.: Imputation method for lifetime exposure assessment in air pollution epidemiologic studies. *Environmental Health* 2013 **12**:62.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

